Equations of States in Statistical Learning for a Nonparametrizable and Regular Case

Sumio Watanabe
PI Lab., Tokyo Institute of Technology,
Mailbox R2-5, 4259, Midori-ku, Yokohama, 226-8503 Japan
(E-mail) swatanab(AT)pi.titech.ac.jp

June 3, 2009

Abstract

Many learning machines that have hierarchical structure or hidden variables are now being used in information science, artificial intelligence, and bioinformatics. However, several learning machines used in such fields are not regular but singular statistical models, hence their generalization performance is still left unknown. To overcome these problems, in the previous papers, we proved new equations in statistical learning, by which we can estimate the Bayes generalization loss from the Bayes training loss and the functional variance, on the condition that the true distribution is a singularity contained in a learning machine. In this paper, we prove that the same equations hold even if a true distribution is not contained in a parametric model. Also we prove that, the proposed equations in a regular case are asymptotically equivalent to the Takeuchi information criterion. Therefore, the proposed equations are always applicable without any condition on the unknown true distribution.

1 Introduction

Nowadays, a lot of learning machines are being used in information science, artificial intelligence, and bioinformatics. However, several learning machines used in such fields, for example, three-layer neural networks, hidden Markov models, normal mixtures, binomial mixtures, Boltzmann machines, and reduced rank regressions have hierarchical structure or hidden variables, with the result that the mapping from the parameter to the probability distribution is not one-to-one. In such learning machines, it was pointed out that the maximum likelihood estimator is not subject to the normal distribution [5, 4, 6, 2], and that the *a posteriori* distribution can not be approximated by any gaussian distribution [11, 13, 14, 15]. Hence the conventional statistical methods for model selection, hypothesis test, and hyperparameter optimization are not applicable to such learning machines. In other words, we have not yet established the theoretical foundation for learning machines which extract hidden structures from random samples.

In statistical learning theory, we study the problem of learning and generalization based on several assumptions. Let $q(x)$ be a true probability density function and $p(x|w)$ be a learning machine, which is represented by a probability density function

of x for a parameter w. In this paper, we examine the following two assumptions.
(1) The first is the parametrizability condition. A true distribution $q(x)$ is said to be *parametrizable* by a learning machine $p(x|w)$, if there is a parameter w_0 which satisfies $q(x) = p(x|w_0)$. If otherwise, it is called *nonparametrizable*.
(2) The second is the regularity condition. A true distribution $q(x)$ is said to be *regular* for a learning machine $p(x|w)$, if the parameter w_0 that minimizes the log loss function

$$L(w) = -\int q(x) \log p(x|w) dx \tag{1}$$

is unique and if the Hessian matrix $\nabla^2 L(w_0)$ is positive definite. If a true distribution is not regular for a learning machine, then it is said to be *singular*.

In study of layered neural networks and normal mixtures, both conditions are important. In fact, if a learning machine is redundant compared to a true distribution, then the true distribution is parametrizable and singular. Or if a learning machine is too simple to approximate a true distribution, then the true distribution is nonparametrizable and regular. In practical applications, we need a method to determine the optimal learning machine, therefore, a general formula is desirable by which the generalization loss can be estimated from the training loss without regard to such conditions.

In the previous papers [18, 19, 20, 21, 22], we studied a case when a true distribution is parametrizable and singular, and proved new formulas which enable us to estimate the generalization loss from the training loss and the functional variance. Since the new formulas hold for an arbitrary set of a true distribution, a learning machine, and an *a priori* distribution, they are called *equations of states in statistical estimation*. However, it has not been clarified whether they hold or not in a nonparametrizable case.

In this paper, we study the case when a true distribution is nonparametrizable and regular, and prove that the same equations of states also hold. Moreover, we show that, in a nonparametrizable and regular case, the equations of states are asymptotically equivalent to the Takeuchi information criterion (TIC) for the maximum likelihood method. Here TIC was derived for the model selection criterion in the case when the true distribution is not contained in a statistical model [10]. The network information criterion [7] was devised by generalizing it to an arbitrary loss function in the regular case.

If a true distribution is singular for a learning machine, TIC is ill-defined, whereas the equations of states are well-defined and equal to the average generalization losses. Therefore, equations of states can be understood as the generalized version of TIC from the maximum likelihood method in a regular case to Bayes method for regular and singular cases.

This paper consists of six sections. In Section 2, we summarized the framework of

Bayes learning and the results of previous papers. In Section 3, we show the main results of this paper. In Section 4, some lemmas are prepared which are used in the proofs of the main results. The proofs of lemmas are given in the Appendix. In Section 5, we prove the main theorems. In Section 5 and 6, we discuss and conclude this paper.

2 Background

In this section, we summarize the background of the paper.

2.1 Bayes learning

Firstly we introduce the framework of Bayes and Gibbs estimations, which is well known in statistics and learning theory.

Let N be a natural number and $\mathbf{R}^N$ be the N-dimensional Euclidean space. Assume that an information source is given by a probability density function $q(x)$ on $\mathbf{R}^N$ and that random samples $X_1, X_2, ..., X_n$ are independently subject to the probability distribution $q(x)dx$. Sometimes $X_1, X_2, .., X_n$ are said to be training samples and the information source $q(x)$ is called a true probability density function. In this paper we use notations for a given function $g(x)$,

$$\begin{aligned} E_X[g(X)] &= \int g(x)q(x)dx, \\ E_j^{(n)}[g(X_j)] &= \frac{1}{n}\sum_{j=1}^{n} g(X_j). \end{aligned}$$

Note that the expectation $E_X[g(X)]$ is given by the integration by the true distribution, but that the empirical expectation $E_j^{(n)}[g(X_j)]$ can be calculated using random samples.

We study a learning machine $p(x|w)$ of $x \in \mathbf{R}^N$ for a given parameter $w \in \mathbf{R}^d$. Let $\varphi(w)$ be an *a priori* probability density function on $\mathbf{R}^d$. The expectation operator $E_w[\ ]$ by the *a posteriori* probability distribution with the inverse temperature $\beta > 0$ for a given function $g(w)$ is defined by

$$E_w[g(w)] = \frac{1}{Z(\beta)} \int g(w)\varphi(w) \prod_{i=1}^{n} p(X_i|w)^{\beta}\, dw,$$

where $Z(\beta)$ is the normalizing constant. The Bayes generalization loss B_g, the Bayes training loss B_t, the Gibbs generalization loss G_g, and the Gibbs training loss G_t are respectively defined by

$$\begin{aligned} B_g &= -E_X[\,\log E_w[p(X|w)]\,], \\ B_t &= -E_j^{(n)}[\,\log E_w[p(X_j|w)]\,], \\ G_g &= -E_X[\,E_w[\log p(X|w)]\,], \\ G_t &= -E_j^{(n)}[\,E_w[\log p(X_j|w)]\,]. \end{aligned}$$

The functional variance V is defined by

$$V = n \times E_j^{(n)} \{ E_w[\ (\log p(X_j|w))^2\ ] - E_w[\log p(X_j|w)]^2 \}.$$

The concept of the functional variance was firstly proposed in the papers [18, 19, 20, 21]. In this paper, we show that the functional variance plays an important role in learning theory. Remark that B_g, B_t, G_g, G_t, and V are random variables because $E_w[\ \ ]$ depends on random samples. Let $E[\ \ ]$ denote the expectation value overall sets of training samples. Then $E[B_g]$ and $E[B_t]$ are respectively called the average Bayes generalization and training error, and $E[G_g]$ and $E[G_t]$ the average Gibbs ones.

In theoretical analysis, we assume some conditions on a true distribution and a learning machine. If there exists a parameter w_0 such that $q(x) = p(x|w_0)$, then the true distribution is said to be parametrizable. If otherwise, nonparametrizable. In both cases, we define w_0 as the parameter that minimizes the log loss function $L(w)$ in eq.(1). Note that w_0 is equal to the parameter that minimizes the Kullback-Leibler distance from the true distribution to the parametric model. If w_0 is unique and if the Hessian matrix

$$\frac{\partial^2}{\partial w_j \partial w_k} L(w_0)$$

is positive definite, then the true distribution is said to be regular for a learning machine.

Remark. Several learning machines such as a layered neural network or a normal mixture have natural nonidentifiability by the symmetry of a parameter. For example, in a normal mixture,

$$p(x|a,b,c) = \frac{a}{\sqrt{2\pi}}\, e^{-|x-b|^2/2} + \frac{1-a}{\sqrt{2\pi}}\, e^{-|x-c|^2/2},$$

two probability distributions $p(x|a,b,c)$ and $p(1-a,c,b)$ give the same probability distribution, hence the parameter w_0 that minimizes $L(w)$ is not unique for any true distribution. In a parametrizable and singular case, such nonidentifiability strongly affects learning [11, 13]. However, in a nonparametrizable and regular case, the *a posteriori* distribution in the neighborhood of each optimal parameter has the same form, resulting that we can assume w_0 is unique without loss of generality.

2.2 Notations

Secondly, we explain some notations.

For given scalar functions $f(w)$ and $g(w)$, the vector $\nabla f(w)$ and two matrices

$\nabla f(w)\nabla g(w)$ and $\nabla^2 f(w)$ are respectively defined by

$$\begin{aligned}(\nabla f(w))_j &= \frac{\partial f(w)}{\partial w_j},\\ (\nabla f(w)\nabla g(w))_{jk} &= \frac{\partial f(w)}{\partial w_j}\frac{\partial g(w)}{\partial w_k},\\ (\nabla^2 f(w))_{jk} &= \frac{\partial^2 f(w)}{\partial w_j \partial w_k}.\end{aligned}$$

Let n be the number of training samples. For a given constant α, we use the following notations.

(1) $Y_n = O_p(n^\alpha)$ shows that a random variable Y_n satisfies $|Y_n| \leq Cn^\alpha$ with some random variable $C \geq 0$.

(2) $Y_n = o_p(n^\alpha)$ shows that a random variable Y_n satisfies the convergence in probability $|Y_n|/n^\alpha \to 0$.

(3) $y_n = O(n^\alpha)$ shows that a sequence y_n satisfies $|y_n| \leq Cn^\alpha$ with some constant $C \geq 0$.

(4) $y_n = o(n^\alpha)$ shows that a sequence y_n satisfies the convergence $|y_n|/n^\alpha \to 0$.

Remark. For a sequence of random variables, it needs mathematically technical procedure to prove convergence in probability or convergence in law. If we adopt the completely mathematical procedure in the proof, a lot of readers in information science may not find the essential points in the theorems. For example, see [18, 21, 22]. Therefore, in this paper, we adopt the natural and appropriate level of mathematical rigorousness, from the viewpoint of mathematical sciences. The notations O_p and o_p are very useful and understandable for such a purpose.

2.3 Parametrizable and singular case

Thirdly, we introduce the results of the previous researches [18, 19, 20, 21]. We do not prove these results in this paper.

Assume that a true distribution is parametrizable. Even if the true distribution is singular for a learning machine,

$$E[B_g] = S_0 + \frac{\lambda_0 - \nu_0}{n\beta} + \frac{\nu_0}{n} + o(\frac{1}{n}), \tag{2}$$

$$E[B_t] = S_0 + \frac{\lambda_0 - \nu_0}{n\beta} - \frac{\nu_0}{n} + o(\frac{1}{n}), \tag{3}$$

$$E[G_g] = S_0 + \frac{\lambda_0}{n\beta} + \frac{\nu_0}{n} + o(\frac{1}{n}), \tag{4}$$

$$E[G_t] = S_0 + \frac{\lambda_0}{n\beta} - \frac{\nu_0}{n} + o(\frac{1}{n}), \tag{5}$$

$$E[V] = \frac{2\nu_0}{\beta} + o(1), \tag{6}$$

where S_0 is the entropy of the true probability density function $q(x)$,

$$S_0 = -\int q(x) \log q(x) dx.$$

The constants λ_0 and ν_0 are respectively the generalized log canonical threshold and the singular fluctuation, which are birational invariants. The concrete values of them can be derived by using algebraic geometrical transformation called resolution of singularities. By elliminating λ_0 and ν_0 from eq.(2)-eq.(6),

$$E[B_g] = E[B_t] + (\beta/n)E[V] + o(\frac{1}{n}), \tag{7}$$

$$E[G_g] = E[G_t] + (\beta/n)E[V] + o(\frac{1}{n}), \tag{8}$$

hold, which are called *equations of states in learning*, because these relations hold for an arbitrary set of a true distribution, a learning machine, and an *a priori* distribution. By this relation, we can estimate the generalization loss using the training loss and the functional variance. However, it has been left unknown whether the equations of states, eq.(7) and eq.(8), hold or not in nonparametrizable cases.

3 Main Results

In this section, we describe the main results of this paper. The proofs of theorems are given in Section 5.

3.1 Equations of states

In this paper, study the case when a true distribution is nonparametrizable and regular. Three constants S, λ, and ν are respectively defined by the following equations. Let w_0 be the unique parameter that minimizes $L(w)$. Three constants are defined by

$$S = L(w_0), \tag{9}$$

$$\lambda = \frac{d}{2}, \tag{10}$$

$$\nu = \frac{1}{2}\mathrm{tr}(IJ^{-1}), \tag{11}$$

where d is the dimension of the parameter, and I and J are $d \times d$ matrices defined by

$$I = \int \nabla \log p(x|w_0) \nabla \log p(x|w_0) q(x) dx, \tag{12}$$

$$J = -\int \nabla^2 \log p(x|w_0) q(x) dx. \tag{13}$$

Theorem 1 *Assume that a true distribution $q(x)$ is nonparametrizable and regular for a learning machine $p(x|w)$. Then*

$$\begin{aligned}
E[B_g] &= S + \frac{\lambda - \nu}{n\beta} + \frac{\nu}{n} + o(\frac{1}{n}), && (14)\\
E[B_t] &= S + \frac{\lambda - \nu}{n\beta} - \frac{\nu}{n} + o(\frac{1}{n}), && (15)\\
E[G_g] &= S + \frac{\lambda}{n\beta} + \frac{\nu}{n} + o(\frac{1}{n}), && (16)\\
E[G_t] &= S + \frac{\lambda}{n\beta} - \frac{\nu}{n} + o(\frac{1}{n}), && (17)\\
E[V] &= \frac{2\nu}{\beta} + o(1). && (18)
\end{aligned}$$

Therefore, equations of states hold,

$$\begin{aligned}
E[B_g] &= E[B_t] + (\beta/n)E[V] + o(\frac{1}{n}), && (19)\\
E[G_g] &= E[G_t] + (\beta/n)E[V] + o(\frac{1}{n}). && (20)
\end{aligned}$$

Proof of this theorem is given in Section 5. Note that constants are different between the parametrizable and nonparametrizable cases, that is to say, $S \neq S_0$, $\lambda \neq \lambda_0$, and $\nu \neq \nu_0$. However, the same equations of states still hold. In fact, eq.(19) and eq.(20) are completely equal to as eq.(7) and eq.(8), respectively.

By combining the results of the previous papers with the new result in Theorem 1, it is ensured that the equations of states are applicable to arbitrary set of a true distribution, a learning machine, and an *a priori* distribution, without regard to the condition on the unknown true distribution.

Remark. If a true distribution is parametrizable and regular, then $I = J$, hence $\lambda = \nu = d/2$. If otherwise, $I \neq J$ in general. Note that J is positive definite by the assumption, but that I may not be positive definite in general.

3.2 Comparison TIC with equations of states

If the maximum likelihood method is employed, or equivalently if $\beta = \infty$, then B_g and B_t are respectively equal to the generalization and training losses of the maximum likelihood method. It was proved in [10] that

$$E[B_g] = E[B_t] + \frac{TIC}{n} + o(\frac{1}{n}) \quad (\beta = \infty), \tag{21}$$

where

$$TIC = \mathrm{tr}(I(w_0)J(w_0)^{-1}).$$

On the other hand, the equations of states, eq.(19) in Theorem 1 show that,

$$E[B_g] = E[B_t] + \frac{E[\beta V]}{n} + o(\frac{1}{n}). \quad (0 < \beta < \infty), \tag{22}$$

Therefore, in this subsection, let us compare βV with TIC in the nonparametrizable and regular case.

Let $L_n(w)$ be the empirical log loss function

$$L_n(w) = -E_j^{(n)}[\log p(X_j|w)] - \frac{1}{n\beta}\log \varphi(w).$$

Three matrices are defined by

$$\begin{aligned} I_n(w) &= E_j^{(n)}[\nabla \log p(X_j|w) \nabla \log p(X_j|w)], &(23)\\ J_n(w) &= -E_j^{(n)}[\nabla^2 \log p(X_j|w)], &(24)\\ K_n(w) &= \nabla^2 L_n(w). &(25) \end{aligned}$$

In practical applications, instead of TIC, the empirical TIC is employed,

$$TIC_n = \mathrm{tr}(I_n(w_{MLE}) J_n(w_{MLE})^{-1}),$$

where w_{MLE} is the maximum likelihood estimator. Then by using the convergence in probability $w_{MLE} \to w_0$,

$$E[TIC_n] = TIC + o(1).$$

On the other hand, we have shown in Theorem 1,

$$E[\beta V] = TIC + o(1).$$

Hence let us compare βV with TIC_n as random variables.

Theorem 2 *Assume that $q(x)$ is nonparametrizable and regular for a learning machine $p(x|w)$. Then*

$$\begin{aligned} TIC_n &= TIC + O_p(\frac{1}{\sqrt{n}}), \\ \beta V &= TIC + O_p(\frac{1}{\sqrt{n}}), \\ \beta V &= TIC_n + O_p(\frac{1}{n}). \end{aligned}$$

Proof of this theorem is given in Section 5. Theorem 2 shows that the difference between βV and TIC_n is in the smaller order than the variance of them. Therefore, if a true distribution is nonparametrizable and regular for a learning machine, then the equations of states are asymptotically equivalent to the empirical TIC. If a true distribution is singular or if the number of training samples are not so large, then the empirical TIC and the equations of states are not equivalent, in general. Hence the equations of states are applicable more widely than TIC. Experimental analysis for the equations of states was reported in [18, 19, 20]. The main purpose of this paper is to prove Theorems 1 and 2. Its application to practical problems is a topic for the future study.

4 Preparation of Proof

In this section, we summarize the basic properties which are used in the proofs of main theorems.

4.1 Maximum *a posteriori* estimator

Firstly, we study the asymptotic behavior of the maximum *a posteriori* estimator. By the definition, for each w,

$$K_n(w) = J_n(w) + O(\frac{1}{n}).$$

By the central limit theorem, for each w,

$$I_n(w) = I(w) + O_p(\frac{1}{\sqrt{n}}), \tag{26}$$

$$J_n(w) = J(w) + O_p(\frac{1}{\sqrt{n}}), \tag{27}$$

$$K_n(w) = J(w) + O_p(\frac{1}{\sqrt{n}}). \tag{28}$$

The parameter that minimizes $L_n(w)$ is denoted by $\hat{w}$, which is called the maximum *a posteriori* estimator (MAP). If $\beta = 1$, then it is equal to the conventional maximum *a posteriori* estimator (MAP). If $\beta = \infty$, or equivalently $1/\beta = 0$, then it is the maximum likelihood estimator (MLE), which is denoted by w_{MLE}.

Let us summarize the basic properties of the maximum *a posteriori* estimator. Becaue w_0 and $\hat{w}$ minimizes $L(w)$ and $L_n(w)$ respectively,

$$\nabla L(w_0) = 0, \tag{29}$$

$$\nabla L_n(\hat{w}) = 0. \tag{30}$$

By the assumption, w_0 is unique and the matrix J is positive definite, the consistency of $\hat{w}$ holds under the natural condition, in other words, the convergences in probability $\hat{w} \to w_0$ $(n \to \infty)$ hold for $0 < \beta \leq \infty$. In this paper, we assume such consistency condition.

From eq.(30), there exists w^*_β which satisfies

$$\nabla L_n(w_0) + \nabla^2 L_n(w^*_\beta)(\hat{w} - w_0) = 0 \tag{31}$$

and

$$|w^*_\beta - w_0| \leq |\hat{w} - w_0|,$$

where $|\cdot|$ denotes the norm of $\mathbf{R}^d$. By using the definition $K_n(w^*_\beta) = \nabla^2 L_n(w^*_\beta)$,

$$\hat{w} - w_0 = -K_n(w^*_\beta)^{-1} \nabla L_n(w_0). \tag{32}$$

By using the law of large numbers and the central limit theorem, $K_n(w^*_\beta)$ converges to J in probability and $\sqrt{n}\,\nabla L_n(w_0)$ converges in law to the normal distribution with average 0 and covariance matrix I. Therefore

$$\sqrt{n}\;(\hat{w}-w_0)$$

converges in law to the normal distribution with average 0 and covariance matrix $J^{-1}IJ^{-1}$, resulting that

$$E[(\hat{w}-w_0)(\hat{w}-w_0)^T] = \frac{J^{-1}IJ^{-1}}{n} + o(\frac{1}{n}), \tag{33}$$

for $0 < \beta \leq \infty$, where $(\ \ )^T$ denotes the transposed vector. In other words,

$$\hat{w} = w_0 + O_p(\frac{1}{\sqrt{n}}). \tag{34}$$

Hence,

$$K_n(w^*_\beta) = J(w_0) + O_p(\frac{1}{\sqrt{n}}).$$

By using eq.(32),

$$\hat{w} - w_{MLE} = \Big(K_n(w^*_\infty)^{-1} - K_n(w^*_\beta)^{-1}\Big)\nabla L_n(w_0).$$

Since $\nabla L_n(w_0) = O_p(1/\sqrt{n})$ and $J(w_0)$ is positive definite, we have

$$w_{MLE} = \hat{w} + O_p(\frac{1}{n}). \tag{35}$$

4.2 Expectations by *a posteriori* distribution

Secondly, the behavior of the *a posteriori* distribution is described as follows.

For a given function $g(w)$, the average by the *a posteriori* distribution is defined by

$$E_w[g(w)] = \frac{\int g(w)\exp(-n\beta L_n(w))dw}{\int \exp(-n\beta L_n(w))dw}.$$

Then we can prove the following relations.

Lemma 1

$$\begin{aligned}
E_w[(w-\hat{w})] &= O_p(\frac{1}{n}), & (36)\\
E_w[(w-\hat{w})(w-\hat{w})^T] &= \frac{K_n(\hat{w})^{-1}}{n\beta} + O_p(\frac{1}{n^2}), & (37)\\
E_w[(w-\hat{w})_i(w-\hat{w})_j(w-\hat{w})_k] &= O_p(\frac{1}{n^2}), & (38)\\
E_w[|w-\hat{w}|^m] &= O_p(\frac{1}{n^{m/2}}) \quad (m \geq 1). & (39)
\end{aligned}$$

Moreover,

$$EE_w[(w-w_0)(w-w_0)^T] = \frac{J^{-1}IJ^{-1}}{n} + \frac{J^{-1}}{n\beta} + o(\frac{1}{n}). \quad (40)$$

$$EE_w[|w-w_0|^3] = o(\frac{1}{n}). \quad (41)$$

For the proof of this lemma, see Appendix.

Let us introduce a log density ratio function $f(x,w)$ by

$$f(x,w) = \log \frac{p(x|w_0)}{p(x|w)}.$$

Then $f(x,w_0) \equiv 0$ and

$$\begin{aligned} \nabla f(x,w) &= -\nabla \log p(x|w), \\ \nabla^2 f(x,w) &= -\nabla^2 \log p(x|w). \end{aligned}$$

In the proof of Theorems 1, we need the following six expectation values,

$$\begin{aligned} D_1 &= EE_X[E_w[f(X,w)]], \\ D_2 &= (1/2)EE_X[E_w[\ f(X,w)^2\ ]], \\ D_3 &= (1/2)EE_X[\ E_w[f(X,w)]^2\ ], \\ D_4 &= EE_j^{(n)}[E_w[f(X_j,w)]], \\ D_5 &= (1/2)EE_j^{(n)}[E_w[\ f(X_j,w)^2\ ]], \\ D_6 &= (1/2)EE_j^{(n)}[\ E_w[f(X_j,w)]^2\ ]. \end{aligned}$$

The constant μ is defined by

$$\mu = \frac{1}{2}\mathrm{tr}(IJ^{-1}IJ^{-1}). \quad (42)$$

Then we can prove the following relations.

Lemma 2 *Let ν and μ be constants which are respectively defined by eq.(11) and eq.(42). Then*

$$\begin{aligned} D_1 &= \frac{d}{2n\beta} + \frac{\nu}{n} + o(\frac{1}{n}), \\ D_2 &= \frac{\nu}{n\beta} + \frac{\mu}{n} + o(\frac{1}{n}), \\ D_3 &= \frac{\mu}{n} + o(\frac{1}{n}), \\ D_4 &= \frac{d}{2n\beta} - \frac{\nu}{n} + o(\frac{1}{n}), \\ D_5 &= \frac{\nu}{n\beta} + \frac{\mu}{n} + o(\frac{1}{n}), \\ D_6 &= \frac{\mu}{n} + o(\frac{1}{n}). \end{aligned}$$

For the proof of this lemma, see Appendix.

5 Proofs

In this section, we prove theorems.

5.1 Proof of Theorem 1

Firstly, by using the definitions

$$\begin{aligned} S &= L(w_0) = -E_X[\log p(X|w_0)], \\ p(x|w) &= p(x|w_0)\exp(-f(x,w)), \end{aligned}$$

the Bayes generalization loss is given by

$$\begin{aligned} E[B_g] &= -EE_X \log E_w[p(X|w)] \\ &= S - EE_X[\log E_w[\exp(-f(X,w))]] \\ &= S - EE_X[\log E_w(1 - f(X,w) + \frac{f(X,w)^2}{2})] + o(\frac{1}{n}) \\ &= S + EE_X E_w[f(X,w)] - \frac{1}{2}EE_X E_w[f(X,w)^2] \\ &\quad + \frac{1}{2}EE_X[\ E_w[f(X,w)]^2\ ] + o(\frac{1}{n}) \\ &= S + D_1 - D_2 + D_3 + o(\frac{1}{n}) \\ &= S + \frac{d}{2n\beta} - \frac{\nu}{n\beta} + \frac{\nu}{n} + o(\frac{1}{n}). \end{aligned}$$

Secondly, the Bayes training loss is

$$\begin{aligned} E[B_t] &= -EE_j^{(n)} \log E_w[p(X_j|w)] \\ &= S - EE_j^{(n)}[\log E_w[\exp(-f(X_j,w))]] \\ &= S - EE_j^{(n)}[\log E_w(1 - f(X_j,w) + \frac{f(X_j,w)^2}{2})] + o(\frac{1}{n}) \\ &= S + EE_j^{(n)} E_w[f(X_j,w)] - \frac{1}{2}EE_j^{(n)} E_w[f(X_j,w)^2] \\ &\quad + \frac{1}{2}EE_j^{(n)}[\ E_w[f(X_j,w)]^2\ ] + o(\frac{1}{n}) \\ &= S + D_4 - D_5 + D_6 + o(\frac{1}{n}) \\ &= S + \frac{d}{2n\beta} - \frac{\nu}{n\beta} - \frac{\nu}{n} + o(\frac{1}{n}). \end{aligned}$$

Thirdly, the Gibbs generalization loss is

$$\begin{aligned} E[G_g] &= -EE_X E_w[\log p(X|w)] \\ &= S + EE_X E_w[f(X,w)] \\ &= S + D_1 \\ &= S + \frac{d}{2n\beta} + \frac{\nu}{n} + o(\frac{1}{n}). \end{aligned}$$

Forthly, the Gibbs training loss is

$$\begin{aligned} E[G_t] &= -EE_X^{(n)}E_w[\log p(X|w)] \\ &= S + EE_X^{(n)}E_w[f(X,w)] \\ &= S + D_4 \\ &= S + \frac{d}{2n\beta} - \frac{\nu}{n} + o(\frac{1}{n}). \end{aligned}$$

Lastly, the functional variance is given by

$$\begin{aligned} E[V] &= 2n(D_5 - D_6) \\ &= 2n(D_2 - D_3) + o(1) \\ &= \frac{2\nu}{\beta} + o(1). \end{aligned}$$

Therefore, we obtained Theorem 1.

5.2 Proof of Theorem 2

Let $V_w[f(X,w)]$ be the variance of $f(X,w)$ in the *a posteriori* distribution,

$$V_w[f(X,w)] \equiv E_w[f(X,w)^2] - E_w[f(X,w)]^2.$$

Then

$$V_w[f(X,w)] = V_w[f(X,w) - f(X,\hat{w})]$$

holds because $f(X,\hat{w})$ is a constant function of w. By the Taylor expansion at $w=\hat{w}$,

$$\begin{aligned} & f(X,w) - f(X,\hat{w}) = \nabla f(X,\hat{w}) \cdot (w-\hat{w}) \\ & + \frac{1}{2}(w-\hat{w}) \cdot \nabla^2 f(X,\hat{w})(w-\hat{w}) + O(|w-\hat{w}|^3). \end{aligned}$$

Using this expansion, and eq.(36), eq.(37), eq.(38), and eq.(39),

$$V_w[f(X,w)] = V_w[(\nabla f(X,\hat{w}))\cdot(w-\hat{w})] + O_p(\frac{1}{n^2}).$$

Hence

$$\begin{aligned} \beta V &\equiv n\beta E_j^{(n)}[V_w[f(X_j,w)]] \\ &= n\beta E_j^{(n)}\{E_w[(\nabla f(X_j,\hat{w})\cdot(w-\hat{w}))^2] \\ &\quad - E_w[\nabla f(X_j,\hat{w})\cdot(w-\hat{w})]^2\} + O_p(\frac{1}{n}). \end{aligned}$$

The second term is $O_p(1/n)$ by eq.(36). Therefore, by applying eq.(37) to the first term,

$$\begin{aligned} \beta V &= n\beta \operatorname{tr}\Big(E_j^{(n)}[(\nabla f(X_j,\hat{w}))(\nabla f(X_j,\hat{w}))^T] \\ &\quad \times E_w[(w-\hat{w})(w-\hat{w})^T]\Big) + O_p(\frac{1}{n}) \\ &= \operatorname{tr}(I_n(\hat{w})K_n^{-1}(\hat{w})) + O_p(\frac{1}{n}). \end{aligned}$$

Therefore, by using eq.(35), proof of Theorem 2 is completed.

6 Discussion

Let us discuss the results of this paper from the three different points of view.

Firstly, we discuss the method how to numerically calculate the equations of states. The widely applicable information criterion (WAIC) [18, 22] is defined by

$$\begin{aligned}\mathrm{WAIC} &= -\sum_{i=1}^{n}\log E_w[p(X_i|w)] \\ &\quad +\beta\sum_{i=1}^{n}\Big\{E_w[(\log p(X_i|w))^2]-E_w[\log p(X_i|w)]^2\Big\}.\end{aligned}$$

Then by Theorem 1,

$$E[WAIC]=E[nB_g]+o(1)$$

holds. Hence by minimization of WAIC, we can optimize the model and the hyperparameter for the minimum Bayes generalization loss. In Bayes estimation, a set of parameters $\{w_k\}$ is prepared so that it approximates the *a posteriori* distribution. Sometimes it is done by the Markov chain Monte Carlo method, and we can approximate the average by the *a posteriori* distribution by

$$E_w[f(w)]\cong\frac{1}{K}\sum_{k=1}^{K}f(w_k).$$

Therefore the WAIC can numerically calculate by such a set $\{w_k\}$.

Secondly, we study the fluctuation of the Bayes generalization error. In Theorem 1, we proved that, as the number of training samples tends to infinity, two expectation values converge to the same value,

$$\begin{aligned}E[n(B_g-B_t)] &\to \mathrm{tr}(IJ^{-1}), \\ E[\beta V] &\to \mathrm{tr}(IJ^{-1}).\end{aligned}$$

Moreover, in Theorem 2, we proved the convergence in probability,

$$\beta V\to\mathrm{tr}(IJ^{-1}).$$

On the other hand, by the same way as Theorem 1, we can prove

$$n(B_g-B_t)=n\times\mathrm{tr}(I(\hat{w}-w_0)(\hat{w}-w_0)^T)+o_p(1).$$

Since $\sqrt{n}(\hat{w}-w_0)$ converges in law to the gaussian random variable whose average is zero and variance is $J^{-1}IJ^{-1}$, the random variable $n(B_g-B_t)$ converges to not a constant in probability but to a random variable in law. In other words, the relation between expectation values

$$E[B_g]=E[B_t]+\frac{\beta E[V]}{n}+o(\frac{1}{n}) \tag{43}$$

holds, whereas they are not equal to each other as random variables,

$$B_g \neq B_t + \frac{\beta V}{n} + o_p(\frac{1}{n}). \tag{44}$$

Note that, even if the true distribution is paramertrizable and regular, the generalization and training losses have same properties, therefore both AIC and TIC have same properties as eq.(43) and eq.(44).

Lastly, let us compare the generalization loss by the Bayes estimation with that by the maximum likelihood estimation. In a regular and parametrizable case, they are equal to each other asymptotically. In a parametrizable and singular case, the Bayes generalization error is smaller than that of the maximum likelihood method. Let us compare them in a nonparametrizable and regular case.

$$E[B_g] = S + \frac{\mathrm{tr}(IJ^{-1})}{2n} + \frac{d - \mathrm{tr}(IJ^{-1})}{2n\beta} + o(\frac{1}{n}).$$

When $\beta = \infty$, this is the generalization error of the maximum likelihood method. If $d > \mathrm{tr}(IJ^{-1})$, then $E[B_g]$ is a decreasing function of $1/\beta$. Or if $d < \mathrm{tr}(IJ^{-1})$, then $E[B_g]$ is an increasing function of $1/\beta$. If $I < J$, then $\mathrm{tr}(IJ^{-1}) < d$. By the definition of I and J,

$$I = \int \nabla p(x|w_0) \nabla p(x|w_0) \frac{q(x)}{p(x|w_0)^2} dx$$

and

$$J = I - Q,$$

where

$$Q = \int (\nabla^2 p(x|w_0)) \ \frac{q(x)}{p(x|w_0)} dx.$$

If $Q < 0$, then $\mathrm{tr}(IJ^{-1}) < d$, resulting that the generalization loss of Bayes estimation is smaller than that by the maximum likelihood method.

Example. For $w \in \mathbf{R}$,

$$p(x|w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-w)^2}{2}),$$

Then

$$L(w) = \frac{1}{2} \int (x-w)^2 q(x) dx + \frac{1}{2} \log(2\pi).$$

Hence $w_0 = E_X[X]$ and

$$L(w_0) = V(X) + \frac{1}{2} \log(2\pi),$$

where $V(X) = E_X[X^2] - E_X[X]^2$. The value Q is

$$Q = V(X) - 1.$$

If $V(X) > 1$, then the generalization error is a decreasing function of $1/\beta$, in other words, the Bayes estimation makes the generalization loss is smaller than that by the maximum likelihood method. Hence, in a nonparametrizable case, it depends on the case which estimation makes the generalization loss smaller.

7 Conclusion

In this paper, we theoretically proved that equations of states in statistical estimation hold even if a true distribution is nonparametrizable and regular for a learning machine. In the previous paper, we proved that the equations of states hold even if a true distribution is parametrizable and singular. By combining these results, the equations of states are applicable without regard to the condition of the true distribution and the learning machine. Moreover, the equations of states contains AIC and TIC in the special cases.

Acknowledgment

This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 18079007.

8 Appendix

8.1 Proof of Lemma 1

By using eq.(30), $\nabla L_n(\hat{w}) = 0$, in a neighborhood of $\hat{w}$,

$$L_n(w) = L_n(\hat{w}) + \frac{1}{2}(w - \hat{w}) \cdot K_n(\hat{w})(w - \hat{w}) + r(w),$$

where $r(w)$ is given by

$$r(w) = \frac{1}{6} \sum_{i,j,k=1}^{d} (\nabla^3 L_n(\hat{w}))_{ijk}(w - \hat{w})_i(w - \hat{w})_j(w - \hat{w})_k + O(|w - \hat{w}|^4).$$

Hence, for a given function $g(w)$, the average by the *a posteriori* distribution is given by

$$E_w[g(w)] \ = \ \frac{\int g(w)\ \exp\Big(-\frac{n\beta}{2}(w-\hat{w}) \cdot K_n(\hat{w})(w-\hat{w}) - n\beta r(w)\Big)dw}{\int \exp\Big(-\frac{n\beta}{2}(w-\hat{w}) \cdot K_n(\hat{w})(w-\hat{w}) - n\beta r(w)\Big)dw}.$$

The main region of the integration is a neighborhood of $\hat{w}$, $|w - \hat{w}| < \epsilon$, hence by putting $w' = \sqrt{n}(w - \hat{w})$,

$$E_w[g(w)] = \frac{\int g(\hat{w} + \frac{w'}{\sqrt{n}})\ \exp(-\frac{\beta}{2}w' \cdot K_n(\hat{w})w' - \frac{\beta\delta(w')}{\sqrt{n}} + O_p(\frac{1}{n}))dw'}{\int \exp(-\frac{\beta}{2}w' \cdot K_n(\hat{w})w' - \frac{\beta\delta(w')}{\sqrt{n}} + O_p(\frac{1}{n}))dw'},$$

where $\delta(w')$ is the third-order polynomial,

$$\delta(w') = \frac{1}{6}\sum_{i,j,k=1}^{d} (\nabla^3 L_n(\hat{w}))_{ijk} w'_i w'_j w'_k.$$

By using

$$\exp\Big(-\frac{\beta\delta(w')}{\sqrt{n}} + O_p(\frac{1}{n})\Big) = 1 - \frac{\beta\delta(w')}{\sqrt{n}} + O_p(\frac{1}{n}),$$

it follows that

$$E_w[g(w)] = \frac{\int g(\hat{w} + \frac{w'}{\sqrt{n}})\ (1 - \frac{\beta\delta(w')}{\sqrt{n}}) \exp(-\frac{\beta}{2} w' \cdot K_n(\hat{w}) w') dw'}{\int \exp(-\frac{\beta}{2} w' \cdot K_n(\hat{w}) w') dw'} + O_p(\frac{1}{n}).$$

Hence by putting $g(w) = w - \hat{w}$, we obtain eq.(36) and by putting $g(w) = (w-\hat{w})(w-\hat{w})^T$, eq.(37). By the same way, eq.(38) and eq.(39) are proved. Let us prove eq.(40). By using eq.(37),

$$\begin{aligned}
& E_w[(w - w_0)(w - w_0)^T] \\
&= E_w[(\hat{w} - w_0 + \frac{w'}{\sqrt{n}})(\hat{w} - w_0 + \frac{w'}{\sqrt{n}})^T] \\
&= (\hat{w} - w_0)(\hat{w} - w_0)^T + \frac{1}{n} E_w[w' w'^T] + O_p(\frac{1}{n^{3/2}}) \\
&= (\hat{w} - w_0)(\hat{w} - w_0)^T + \frac{K_n(\hat{w})^{-1}}{n\beta} + O_p(\frac{1}{n^{3/2}}).
\end{aligned} \tag{45}$$

Then by applying eq.(33), eq.(40) is obtained. Lastly, in general,

$$|w - w_0|^3 \le 3(|w - \hat{w}|^3 + |\hat{w} - w_0|^3).$$

Then, by eq.(34) and eq.(39), eq.(41) is derived. Therefore we have obtained Lemma 1.

8.2 Proof of Lemma 2

By the Taylor expansion of $f(X, w)$ at w_0,

$$\begin{aligned}
f(X, w) &= \nabla f(X, w_0) \cdot (w - w_0) \\
&\quad + \frac{1}{2}(w - w_0) \cdot \nabla^2 f(X|w_0)(w - w_0) \\
&\quad + f_3(X, w),
\end{aligned} \tag{46}$$

where $f_3(X, w)$ satisfies

$$|f_3(X, w)| \le C(X, w)|w - w_0|^3$$

in a neighborhood of w_0 with a function $C(X, w) \ge 0$. Let us estimate $D_1, ..., D_6$. Firstly, by using eq.(29) and eq.(41),

$$\begin{aligned}
D_1 &= \frac{1}{2} E E_w E_X[(w - w_0) \cdot \nabla^2 f(X, w_0)(w - w_0)] + o(\frac{1}{n}) \\
&= \frac{1}{2} E E_w[(w - w_0) \cdot J(w - w_0)] + o(\frac{1}{n}).
\end{aligned}$$

Then by using the identity

$$(\forall u, v \in \mathbf{R}^d, A \in \mathbf{R}^{d\times d}) \quad u \cdot Av = \mathrm{tr}(Avu^T),$$

and eq.(40),

$$\begin{aligned} D_1 &= \frac{1}{2}EE_w[\mathrm{tr}((J(w-w_0))(w-w_0)^T)] + o(\frac{1}{n}) \\ &= \frac{d}{2n\beta} + \frac{\mathrm{tr}(IJ^{-1})}{2n} + o(\frac{1}{n}). \end{aligned}$$

Secondly, by using the identity

$$(\forall u, v \in \mathbf{R}^d) \quad (u \cdot v)^2 = \mathrm{tr}((uu^T)(vv^T)),$$

the definition of I, and eq.(40),

$$\begin{aligned} D_2 &= (1/2)EE_w[\ E_X[(\nabla f(X,w_0)\cdot(w-w_0))^2]\ ] + o(\frac{1}{n}) \\ &= (1/2)EE_w[\mathrm{tr}(I(w-w_0)(w-w_0)^T)] + o(\frac{1}{n}) \\ &= \frac{\mathrm{tr}(IJ^{-1})}{2n\beta} + \frac{\mathrm{tr}(IJ^{-1}IJ^{-1})}{2n} + o(\frac{1}{n}). \end{aligned}$$

Thirdly, by the definition of I, eq.(36), and eq.(33),

$$\begin{aligned} D_3 &= (1/2)EE_X[E_w[\nabla f(X,w_0)\cdot(w-w_0)]^2] + o(\frac{1}{n}) \\ &= (1/2)EE_X[(\nabla f(X,w_0)\cdot(\hat{w}-w_0))^2] + o(\frac{1}{n}) \\ &= (1/2)E[\mathrm{tr}(I(\hat{w}-w_0)(\hat{w}-w_0)^T)] + o(\frac{1}{n}) \\ &= \frac{\mathrm{tr}(IJ^{-1}IJ^{-1})}{2n} + o(\frac{1}{n}). \end{aligned}$$

Fourthly, by the Taylor expansion eq.(46)

$$\begin{aligned} D_4 &= EE_w[E_j^{(n)}[\nabla f(X_j,w_0)\cdot(w-w_0)]] \\ &\quad + \frac{1}{2}EE_w[E_j^{(n)}[(w-w_0)\cdot\nabla^2 f(X_j,w_0)(w-w_0)]] + o(\frac{1}{n}) \\ &= E[\ E_j^{(n)}[\nabla f(X_j,w_0)]\cdot E_w[w-w_0]\ ] \\ &\quad + \frac{1}{2}EE_w[(w-w_0)\cdot J_n(w_0)(w-w_0)]] + o(\frac{1}{n}). \end{aligned}$$

Then by using $E[J_n(w_0)] = J$, the second term is equal to D_1. To the first term, we apply eq.(36) and

$$E_j^{(n)}[\nabla f(X_j,w_0)] = \nabla L_n(w_0) + O_p(\frac{1}{n}),$$

we obtain

$$D_4 = E[(\nabla L_n(w_0))\cdot(\hat{w}-w_0)] + D_1 + o(\frac{1}{n}).$$

Then applying eq.(31), $K_n(w_0) \to J$, and $w^*_\beta \to w_0$,

$$\begin{aligned} D_4 &= -E[(K_n(w^*_\beta)(\hat{w} - w_0)) \cdot (\hat{w} - w_0)] + D_1 + o(\frac{1}{n}) \\ &= \frac{d}{2n\beta} - \frac{\mathrm{tr}(IJ^{-1})}{2n} + o(\frac{1}{n}). \end{aligned}$$

And lastly, by the definitions,

$$\begin{aligned} D_5 &= (1/2)EE_w[\ E^n_j[(\nabla f(X_j, w_0) \cdot (w - w_0))^2]\ ] + o(\frac{1}{n}), \\ D_6 &= (1/2)EE^n_j[E_w[\nabla f(X_j, w_0) \cdot (w - w_0)]^2] + o(\frac{1}{n}). \end{aligned}$$

By using the convergences in probability, $I_n(w_0) \to I$ and $J_n(w_0) \to J$, it follows that

$$\begin{aligned} D_5 &= D_2 + o(\frac{1}{n}), \\ D_6 &= D_3 + o(\frac{1}{n}), \end{aligned}$$

which completes Lemma 2.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, Vol.19, pp.716-723, 1974.

[2] S.-i. Amari, H. Park, and T. Ozeki, “Singularities Affect Dynamics of Learning in Neuromanifolds,” Neural Comput., 18(5),pp.1007 - 1065,2006.

[3] M.Aoyagi, S.Watanabe,“Stochastic complexities of reduced rank regression in Bayesian estimation,” Neural Networks, Vol.18,No.7, pp.924-933, 2005.

[4] K. Hagiwara, “On the Problem in Model Selection of Neural Network Regression in Overrealizable Scenario, ” Neural Comput., Vol.14,Vol.8, pp.1979 - 2002, 2002.

[5] J.A.Hartigan,“A failure of likelihood asymptotics for normal mixture,” Proc. of Barkeley Conf. in honor of Jerzy Neyman and Jack Keifer, Vol.2, pp.807-810,1985.

[6] T. Hayasaka, M. Kitahara, and S. Usui, “On the Asymptotic Distribution of the Least-Squares Estimators in Unidentifiable Models,” Neural Comput., Vol.16 ,No.1, pp.99 - 114, 2004.

[7] N. Murata, S. Yoshizawa, and S. Amari, “Network information criterion - determining the number of hidden units for an artificial neural network model,” IEEE Transactions on Neural Networks, Vol.5, No.6, pp.865-872, November, 1994.

[8] K. Nagata, S. Watanabe, ”Asymptotic Behavior of Exchange Ratio in Exchange Monte Carlo Method”, International Journal of Neural Neworks, to appear.

[9] K. Nagata and S. Watanabe,"Exchange Monte Carlo Sampling from Bayesian Posterior for Singular Learning Machines", IEEE Transactions on Neural Networks, to appear.

[10] K. Takeuchi, "Distribution of statistic and criterion for optimal model selection," Surikagaku, Vol. 153, pp. 12-18, 1976.

[11] S.Watanabe, "Generalized Bayesian framework for neural networks with singular Fisher information matrices," Proc. of International Symposium on Nonlinear Theory and Its applications, (Las Vegas), pp.207-210, 1995.

[12] S.Watanabe, "Algebraic analysis for singular statistical estimation," Proc. of International Journal of Algorithmic Learning Theory, Lecture Notes on Computer Sciences, 1720, pp.39-50, 1999.

[13] S.Watanabe, "Algebraic Analysis for Nonidentifiable Learning Machines," Neural Computation, Vol.13, No.4, pp.899-933, 2001.

[14] S. Watanabe, "Algebraic geometrical methods for hierarchical learning machines," Neural Networks, Vol.14, No.8,pp.1049-1060, 2001.

[15] S. Watanabe, "Learning efficiency of redundant neural networks in Bayesian estimation," IEEE Transactions on Neural Networks, Vol.12, No.6, 1475-1486,2001.

[16] S.Watanabe, S.-I.Amari,"Learning coefficients of layered models when the true distribution mismatches the singularities", Neural Computation, Vol.15,No.5,1013-1033, 2003.

[17] S.Watanabe,"Algebraic geometry of singular learning machines and symmetry of generalization and training errors," Neurocomputing, Vol.67,pp.198-213,2005.

[18] S. Watanabe, "Equations of states in statistical estimation," arXiv:0712.0653, 2007.

[19] S. Watanabe," Generalization and Training Errors of Bayes and Gibbs estimation in Singular Learning Machines," IEICE Technical report, NC2007-75, Vol.2007-12,pp.25-30,2007.

[20] S. Watanabe,"A formula of equations of states in singular learning machines," Proceedings of WCCI, 2008.

[21] S. Watanabe, "A limit theorem in singular regression problem," arXiv:0901.2376v1, 2009.

[22] S. Watanabe, "Algebraic geometry and statistical learning theory," Cambirdge University Press, 2009.

[23] S. Watanabe, "On a relation between a limit theorem in learning theory and singular fluctuation," IEICE Technical Report, NC2008-111, pp.45-50, 2009.

[24] K.Yamazaki, S.Watanabe,"Singularities in mixture models and upper bounds of stochastic complexity." International Journal of Neural Networks, Vol.16, No.7, pp.1029-1038,2003.

[25] K.Yamazaki, S.Watanabe," Singularities in Complete bipartite graph-type Boltzmann machines and upper bounds of stochastic complexities", IEEE Trans. on Neural Networks, Vol. 16 (2), pp.312-324, 2005.

[26] K. Yamazaki and S. Watanabe, "Algebraic geometry and stochastic complexity of hidden Markov models", Neurocomputing, Vol.69, pp.62-84, 2005.